Measuring English Readability for Vietnamese Speakers

Thuan Nguyen and Alexandra L. UITDENBOGERD
RMIT University - School of Science
124 La Trobe St
Melbourne VIC 3000
john.nguyen09@outlook.com,sandra.uitdenbogerd@rmit.edu.au

Abstract

Reading is important for language learners, but text difficulty needs to match a reader's skill level for efficient vocabulary acquisition. Traditional readability measures may not be effective for those who speak English as a second or additional language. This study examines English readability for Vietnamese native speakers (VL1). A collection of text difficulty judgements of nearly 100 English text passages was obtained from 12 VL1 participants, using a 5-point Likert scale. Using features from traditional readability measures, support vector machines and Dale-Chall features gave more accurate predictions than linear models using either Flesch or Dale-Chall features. VL1 participants' text judgements were strongly correlated with their past English test scores. This study introduces a first approximation to readability of English text for VL1, with suggestions for further improvements.

1 Introduction

Extensive reading, that is, reading a large amount of text at a comfortable level of difficulty, is an efficient way to improve language skills, as learners acquire vocabulary as they read, retain it for longer than if they use rote memorisation (Hermann, 2003), and greatly improve their receptive skills (Elley and Mangubhai, 1983). However, the level of difficulty of the text needs to match the learner, for example, to guess the meaning of new words, readers typically need to know at least 95% of the words (Laufer, 1989). Further, text needs to be well below the learner's frustration level (Klare, 1988). As most text written for native (L1) speakers is beyond the beginner and intermediate language learner, students need to start with simple or simplified text and work their way up to more advanced texts as they learn. Thus a method of measuring the readability of text is crucial for language learners, and can be incorporated in to reading recommender systems.

There have been many proposed English text readability measurement techniques, most of which were modelled on native English-speaking (EL1) children, or texts written for them (for example (Flesch, 1948; Dale and Chall, 1948; Schwarm and Ostendorf, 2005)). English comprehension is different for people with different backgrounds and language skills, causing readability measurement techniques developed for EL1 to perform poorly for other L1-L2 combinations (Oller et al., 1972; Uitdenbogerd, 2005). The focus of this study was English readability for Vietnamese speakers (VL1). To our knowledge, no techniques have been built and tested specifically for this cohort.

While there is existing work on readability for non-native (L2) speakers, the majority is trained on data sets that assume texts written for different language levels match the comprehension experienced by L2 speakers. Examples include François and Miltsakaki (2012) for French L2 and Xia et al. (2016) for English. Xia et al. (2016) extended readability measurement techniques by adapting a model trained on texts for English L1 speakers using data from Cambridge English language tests at different Common European Framework of Reference for languages (CEFR) levels. While some studies use corpora that are fairly homogenous, it has been pointed out that other corpora reveal considerable inconsistency across text classes used as ground truth (François, 2014). Furthermore, there is evidence that expert or publisher-based ground truth is a poor surrogate for genuine language learner experience (Vajjala and Lucic, 2019).

This paper replicates two well-known readability measures, Flesch (1948) and Dale and Chall

(1948) by using their features and techniques, but building the model on new data collected from Vietnamese speakers. Prior research has shown that Machine Learning (ML) algorithms and Natural Language Processing (NLP) features provide better results than traditional formulae (François and Miltsakaki, 2012). Therefore, we also tested Support Vector Machines (SVMs) to produce the model for assessing English text readability as perceived by Vietnamese speakers.

Collecting appropriate ground truth was challenging, leading to a smallish data set with a skewed rating distribution. We report on the main techniques relevant to this project (Section 2), details about the ground truth data collection (Section 3), the results of applying linear models and SVMs, and further analysis and discussion of the data set and results.

2 Readability Measurement

Much research has demonstrated that extensive reading increases language acquisition (Hermann, 2003; Elley and Mangubhai, 1983; Laufer, 1989; Klare, 1988). Further research has tried to determine how to select appropriate reading material for learners through the development of readability measurement techniques, either simple metrics that can be applied manually to small samples, or more recently, complex predictive models using NLP features (for example François and Miltsakaki (2012)).

Readability for non-native speakers is likely to be affected by their knowledge of other languages. The principal way that this is experienced is through cognates (and loanwords), that is, words that are similar in appearance and meaning between a pair of languages. Their impact on readability of French for English speakers has been demonstrated (Uitdenbogerd, 2005).

Some studies measure readability (or complexity) of different units of language. While the majority of research estimates readability of whole documents, there is some work on readability of sentences (Pilán et al., 2014) as well as lexical complexity in isolation (Paetzold and Specia, 2016). In this work we look at text passages of 50-200 words in length, being short enough for participants to judge quickly, and long enough to provide context and features.

Flesch (1948) and Dale and Chall (1948) are two of the most popular readability measures for English, both of which extract two statistical features from text and calculate the difficulty level using a linear model. Both measures try to capture the syntactic complexity of a text using the average word count per sentence (WPS). Flesch (1948) captures vocabulary complexity via the average syllable count per word (SPW), whereas Dale and Chall (1948) use a predefined list of 3000 familiar words to calculate the percentage of difficult words (PDW).

The Flesch formula, shown below, produces a readability score that normally falls in the range 0–100 (theoretical maximum would be 121.22), with 0–30 being classed as very difficult, and 90–100 being very easy.

$$\mathrm{RE} = 206.835 - 1.015\left(\frac{\#\text{ wrds}}{\#\text{ snts}}\right) - 84.6\left(\frac{\#\text{ syll}}{\#\text{ wrd}}\right) \quad (1)$$

Dale-Chall's formula calculates the grade-level of text.

$$\mathrm{DC} = 0.0496\left(\frac{\#\text{ words}}{\#\text{ sents}}\right) + 0.1579\left(\frac{\#\text{ hard words}}{\#\text{ words}} * 100\right) \quad (2)$$

Schwarm and Ostendorf (2005) assessed text readability using SVMs and a combination of NLP features and statistical features. The present study also uses SVMs with statistical features from traditional models, but with ground truth data from Vietnamese speakers.

3 Vietnamese Ground truth data collection

Our aim was to collect a text corpus of a wide range of difficulty, and to collect human judgements of their perceived difficulty from VL1 speakers. The intention was to obtain multiple judgements per text to allow some analysis of how different individuals perceive the difficulty of the same text.

We selected a variety of texts to make up ten categories from four different sources: Oxford Bookworms graded readers for learners of English as a second language (EL2) consisting of five levels ranging from level 0 (Starter) to level 4, children's literature, young adult texts, and classic English literature. Oxford Bookworms texts were selected randomly from a digitised data-set. Three children's literature texts were arbitrarily selected from Project Gutenberg's Children's literature bookshelf. Four young adult texts were arbitrarily selected from a library's Young Adult section. The classical literature stories selected were the top three from a top ten list of classics found via a web search. Due to an oversight leading to original classic texts being used instead of the simplified version for Oxford Bookworms levels 5 and 6, there were three times as many texts from classical literature than other sources or levels. Despite the uneven representation of books, with *David Copperfield* and *The Woman in White* having ten samples each, versus only one to five samples for all others, each category was distinct, and had ten randomly selected extracts, allowing sufficient variety for testing readability, and providing a wide range of difficulty.

Extracting sentences from the books was via a script that randomly generated a starting sentence number, and the number of sentences to extract from that point. However, this process was not applied to young adult books since they were not electronically available, therefore a random number generator was used instead, to generate a page number, paragraph number and the number of sentences to extract. Ten texts from each level or source were randomly selected, each around 50-200 words in length, leading to a total of 100 texts. This text length and number of judgements was chosen to minimise the time commitment of volunteer participants and provide sufficient context to assess the readability of the text. It is a similar quantity to samples in previous studies (See for example, Björnsson (1968)). Text was presented to participants in a random order to eliminate ordering effects.

Participants were recruited via an invitation to complete an on-line survey posted in a large Facebook group for Vietnamese students in Australia. Twelve participants completed the entire questionnaire, resulting in 120 samples of data. One participant who did not complete all questions was excluded to avoid potential bias in the data-set toward specific participants' responses. The participants were Vietnamese students studying in Australia, the majority of whom had English IELTS levels 6 to 8, being equivalent to CEFR B2–C1/C2.

Participants were asked to read 10 texts with no time limit, each from a different reading level or source, and to choose an answer based on a 5-point Likert scale, with each point worded specifically for learners of English as a foreign language (Uitdenbogerd et al., 2017) as shown below.

1. The text was very easy. I knew every word.
2. The text was easy, but I did not understand some words.
3. The text was not easy, but I understood the story.
4. The text was difficult. I would need a dictionary.
5. The text was very difficult. A dictionary will not help me.

4 Linear models

Using the ground truth data-set obtained from Vietnamese speakers, we replicated methods used for traditional measures (Flesch, 1948; Dale and Chall, 1948). These new models were built using linear regression and statistical features of texts.

The Natural Language Toolkit (Loper and Bird, 2002) was used to extract the statistical features in the text, and scikit-learn (Pedregosa et al., 2011) was used for training and testing, as well as for calculating mean squared error (MSE). Syllable counts were based on those found in the Carnegie-Mellon Pronouncing Dictionary (cmudict). All words used in the study corpus were in cmudict.

These experiments replicate the techniques and features from Flesch and Dale-Chall, using the collected data-set to feed into the linear regression model. Bootstrapping was also attempted to compensate for the small data size and uneven distribution of responses.

The experiments were set up to train and test ten times on the collected data, the split ratio being 67% and 33% respectively, and each time splitting the data randomly. The aim was to build the model with the least mean squared error (MSE), which measures how well the linear model fits the data; and the least over-fitting amount, measured as the difference between the MSE of the training and test data sets. That is, a good model has a low MSE and a low measure of over-fitting. The coefficients of features of the best run become the recommended model for the given set of features.

When bootstrapping was applied, the data was sampled with replacement 100 times for each Likert scale point, from 1-3. No participants selected 5 (very difficult), and only one selected 4 (difficult), thus 4 was excluded.

4.1 Linear Model based on Flesch features

Table 1 shows the result of using WPS and SPW in the linear regression model based on VL1 judgements of English text difficulty.

The average train MSE and test MSE were 0.25 and 0.31 respectively. The best run produced an MSE of 0.23 for predicting unknown situations, having an over-fitting result of 0.05. The MSE was

Run no	Coeffs (WPS, SPW)	Train MSE	Test MSE	Over-fitting
1	0.005, 1.5	0.29	0.21	0.08
2	0.002, 1.4	0.28	0.23	0.05
3	-0.0004, 0.86	0.30	0.21	0.09
4	0.022, 0.59	0.15	0.51	0.36
5	0.016, 1.355	0.32	0.17	0.15
6	0.008, 1.367	0.30	0.19	0.11
7	0.003, 0.759	0.12	0.57	0.45
8	0.017, 0.459	0.18	0.46	0.28
9	0.013, 0.505	0.35	0.11	0.24
10	0.015, 0.647	0.19	0.41	0.22

Table 1: Results of replicating the Flesch formula

calculated using the following formula:

$$MSE = \frac{\sum_{i=1}^{n}(actual(i) - predicted(i))^2}{n}$$

where n is the number of test samples.

Therefore an average MSE of 0.31 in the test data suggests instability, since the distance between the data rating score and the predicted rating score is the middle of 2 rating levels.

The model can be represented as:

$$\text{VFlesch} = 0.002 * (\frac{\text{\# words}}{\text{\# sents}}) + 1.4 * (\frac{\text{\# sylls}}{\text{\# words}}) \quad (3)$$

This formula shows that the coefficient of SPW has more importance than WPS, at a vocabulary to grammar feature ratio of 700 (See Table 2). This suggests that the original Flesch model, which has a feature ratio of 83.4, has less emphasis on vocabulary, and would therefore be less effective for Vietnamese speakers. On looking at the second and third best runs based on over-fitting score, it can be seen that WPS is consistently small, and in one case is negative, indicating that sentence length can virtually be ignored to get a good estimate of readability for this cohort of speakers. The run with a coefficient ratio most similar to the original Flesch score is Run 5, which has a vocabulary to grammar coefficient ratio of 84.7 and is in the middle of the runs when compared by overfitting amount, indicating that the new coefficients would be more stable.

Applying bootstrapping increased the error significantly with average MSE of 0.54 and resulted in an unpredictable model, so was not helpful in this case.

4.2 Linear Model based on Dale-Chall Features

In Table 3 we report on the model based on WPS and PDW, which are taken from the Dale-Chall formula.

The MSE in predicting training data and test data respectively are 0.28 and 0.24 for the best model, and its coefficients for percentage of difficult words and average word count per sentence are 0.015 and 0.020 respectively. The average MSE across 10 runs are 0.31 and 0.20. This model has less error than the Flesch-based model. The coefficients produced by the runs are also more stable than the Flesch ones, suggesting that the Dale-Chall vocabulary feature is superior. None of the runs produced coefficients with a similar ratio of PDW to WPS as the original Dale-Chall formula (approximately 3.18).

The resulting formula for this model is as follows:

$$\text{VDC} = 0.020(\frac{\text{\# words}}{\text{\# sents}}) + 0.015(\frac{\text{\# hard words}}{\text{\# words}} * 100) \quad (4)$$

In this model, the coefficients of the two features are quite similar to each other. That is, unlike for Flesch, in the Dale-Chall formula WPS is relatively more important for VL1 than vocabulary, since the PDW was weighted 3.18 times more than WPS in the original Dale-Chall formula, but in this model is only 0.75 times (shown in Table 2).

In the previous model, applying bootstrapping increased the error significantly. We also applied bootstrapping in this model to confirm if features are the factor that causes the significant increase in error. Indeed, using bootstrapped data produced a very high level of error ($>$ 0.60 MSE). Therefore, it is safe to conclude that bootstrapping does not work very well with linear models of this data-set.

Additionally, we tested a modified version of the Dale-Chall word list that was potentially more suitable for VL1. The Vietnamese first author of the present study — who found many of the words on the original list unfamiliar and therefore difficult — modified the list by removing any words that seemed difficult. We acknowledge that this is not a robust approach, however it was a good first approximation, and a more representative list for VL1 may be future work. The results of the modified word list were very similar to the original results. Further analysis showed that 26 of the 100 texts contained a slightly higher number of difficult words when using the modified list, 2 texts with 3, 4 with 2 and 20 with 1 respectively, being less than 3% change in a PDW score. Mean (0.35), standard deviation (0.07) and maximum (0.6) PDW remained about the same for both

	Coefficients			Vocab/Grammar Ratios		VL1 Ratio/Orig. Ratio
	WPS	SPW	PDW	SPW/WPS	PDW/WPS	
Coeff. Type	Grammar	Vocab	Vocab			
Original Flesch	1.015	84.6		83.4		
Best VFlesch	0.002	1.4		700		8.4
Original Dale-Chall	0.0496		0.158		3.18	
Best VDC	0.02		0.015		0.75	0.24

Table 2: Vocabulary to grammar coefficient ratios for Flesch, Dale-Chall, and the best linear model runs with VL1 data.

Run no	Coeffs (PDW, WPS)	Train MSE	Test MSE	Over-fitting
1	0.013, 0.014	0.32	0.16	0.16
2	0.010, 0.011	0.34	0.12	0.22
3	0.014, 0.006	0.33	0.16	0.17
4	0.012, 0.020	0.32	0.17	0.15
5	0.014, 0.015	0.25	0.31	0.06
6	0.003, 0.020	0.20	0.40	0.20
7	0.015, 0.020	0.28	0.24	0.04
8	0.009, 0.010	0.35	0.11	0.24
9	0.014, 0.014	0.34	0.13	0.11
10	0.014, 0.013	0.32	0.16	0.16

Table 3: Results of replicating the Dale-Chall formula

versions and the minimum increased from 0.18 to 0.19.

To summarise, the model produced by using features from Dale-Chall gave a lower error rate than the model using Flesch features, and the features appeared to be more stable.

4.3 Combined features from Flesch and Dale-Chall formulas

The features used in this experiment are taken from Flesch and Dale-Chall formulas, which are: WPS, SPW and PDW. Our hypothesis is that this will not affect the model's performance because the Flesch and Dale-Chall formulae try to represent the vocabulary complexity by SPW or PDW respectively, and there may not be much gain by combining the two features. The result confirmed this by producing an MSE of 0.27, which does not provide any improvement on previous models. The two vocabulary features have a correlation of 0.53 for our data-set.

4.4 Using Dale-Chall and Flesch score as a feature

The results of the original Dale-Chall formula and our linear model are in different formats, that is the Dale-Chall score is a grade level ranging from 0-10+ and our model is a difficulty level ranging from 1-5. Therefore, it is not straightforward to scale the result from our model to Dale-Chall and vice versa. Therefore to compare our model with the original Dale-Chall formula we calculated the Dale-Chall score for the text using the original weights and then used it as a feature (Table 4) to calculate the error rate as for previous experiments.

Run no	Coeffs (Dale-Chall)	Train MSE	Test MSE	Over-fitting
1	0.078	0.31	0.19	0.12
2	0.119	0.27	0.26	0.01
3	0.097	0.34	0.13	0.21
4	0.017	0.19	0.44	0.15
5	0.080	0.16	0.49	0.27

Table 4: Results of using Dale-Chall as a feature

The errors of the model fluctuated and produced different results for each random train and test data split resulting high average prediction error across multiple runs. This indicates in some cases the result from Dale-Chall formula does not resemble participant ratings, for example some cases produce a test error of almost 0.50 (run 4 and 5), meaning the model is not learning anything. We conclude that while the *features* in Dale-Chall formula worked best for VL1, the original Dale-Chall formula is less effective for Vietnamese speakers.

When running the same validation against the original Flesch reading ease score, even though the MSE were lower than using original DC as a feature, the same error fluctuation pattern occurs (See Table 5). This suggests that the original Flesch gives a better result than the original DC for VL1, but is still less effective for Vietnamese speakers than than the model trained on DC features with VL1 data.

5 Using SVMs on statistical features

ML is known to be effective in text classification, and has also been applied to text readability (Schwarm and Ostendorf, 2005). Here we ap-

Run no	Coeffs (Flesch)	Train MSE	Test MSE	Over-fitting
1	-0.013	0.31	0.15	0.24
2	-0.008	0.15	0.49	0.34
3	-0.011	0.30	0.19	0.11
4	-0.012	0.29	0.20	0.04
5	-0.009	0.23	0.31	0.14

Table 5: Results of using Flesch as a feature

ply SVMs with a radial basis kernel function to determine whether they will improve readability assessment of English for VL1.

We tested three feature sets: Flesch only, Dale-Chall only, and the combined features of both. The split ratio was different for this experiment, being 70% training data and 30% test data due to the small data-set. As the data-set was quite small for SVMs to be effective we also used bootstrapping to re-sample the data-set from 120 samples to 300 samples, despite the method increasing the error rate in the previous experiments.

We used cross-validation, with the training and test data split randomly for each cross-validation. The experiment was to start with 5 runs and increased to 10 runs for each data-set. The aim of this was to observe the MSE and the variance to see if more cross-validation increases variance, which might indicate an unreliable model. The reason 5-10 runs were chosen is that for each run we generated a random cross-validation to train and test, and since the data is small, after 10 runs it is possible that the model will be over-trained and produce an unreliable prediction model.

The results of using raw data are shown in Table 6 and Table 7.

Feature set	MSE (+/- var.)
Flesch Features	0.14 (+/- 0.08)
Dale-Chall Features	0.15 (+/- 0.09)
Combined features	0.15 (+/- 0.09)

Table 6: 5 runs of SVMs on raw data

Feature set	MSE (+/- var.)
Flesch Features	0.19 (+/- 0.13)
Dale-Chall Features	0.19 (+/- 0.13)
Combined features	0.19 (+/- 0.13)

Table 7: 10 runs of SVMs on raw data

Observing the results, there are two things to notice. Firstly, the MSE increases significantly and the variance also increases, which indicates that the model becomes over-fitted to the data and increased validation increases the error. This can be caused by the small data-set since ML requires large data-sets to be effective.

The data was then re-sampled using the bootstrap method to increase to 300 samples, even though 300 is not a large number for ML (previous ML experiments all have roughly more than 1000 data-points (François and Miltsakaki, 2012; Schwarm and Ostendorf, 2005)), but since it is re-sampled from 120 data-points, then 300 is a reasonable number. The results are shown in Table 8 and Table 9.

Feature set	MSE (+/- var.)
Flesch Features	0.39 (+/- 0.07)
Dale-Chall Features	0.21 (+/- 0.05)
Combined features	0.28 (+/- 0.03)

Table 8: Five runs of SVMs on bootstrapped data

Feature set	MSE (+/- var.)
Flesch Features	0.38 (+/- 0.07)
Dale-Chall Features	0.21 (+/- 0.06)
Combined features	0.27 (+/- 0.07)

Table 9: Ten runs of SVMs on bootstrapped data

In this experiment with the bootstrap method, the results stay almost consistent, even with more cross-validation. This model is confirmed to be more effective, since ML requires a large data-set. The model also gives a better performance than the linear models because it isn't prone to over-fitting, even though the bootstrap method increases MSE.

Additionally, features from the Dale-Chall formula have the best performance across the three feature sets tested. Features from Flesch produced the worst performance, being even worse than linear models, while combined features were not far behind and were comparable to results of linear models.

6 Analysis

In this section we examine some properties of the judgements that were collected, including English skill level of participants, and how that related to their text ratings. We also examine general properties of the texts in each category in the hope of shedding some light on the slightly contradictory results occurring in the readability models.

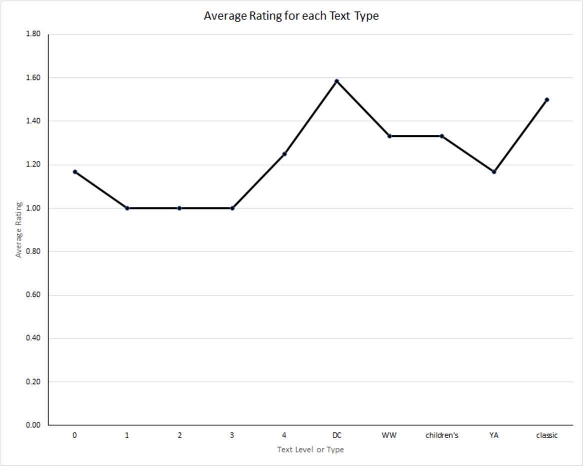

Figure 1: Average rating over twelve human judgements for each type of text.

6.1 Judgements

We collected 10 judgements for each category of text, each from a different participant. Due to the method of random allocation of texts to participants, not all texts in the collection received judgements and some texts received up to four judgements. There were 6–8 text samples judged from each category. Figure 1 shows the average judgements for each text category. When a regression line is fit between the five Oxford Bookworms levels and their averages, the resulting equation had a slope of 0.017 and an R^2 of 0.05, suggesting a very poor fit. This is likely because all the Bookworms texts were too easy for the pool of participants, leading to insufficient difference in ratings across the levels. It can be observed that levels 1-3 all had the same average rating of 1. That is, every participant rated all texts of those levels as being very easy. For level 0, one participant gave a rating of 3 to a text (10). The same participant was the only one to rate any text with a 4 (difficult), and had the highest average ratings of difficulty. Two participants rated all texts as very easy, and therefore provided no information to the models of readability.

Nine of the twelve participants had provided their past IELTS test score. We compared their average ratings and their past IELTS test score (See Figure 2). The R^2 was 0.779, thus a high correlation (0.88). The rating distribution, however, was exponential, with 99, 19, 3, 1 and 0 ratings respectively from very easy to very difficult (fitting a line to the log of the non-zero rating counts has an R^2 of 0.99).

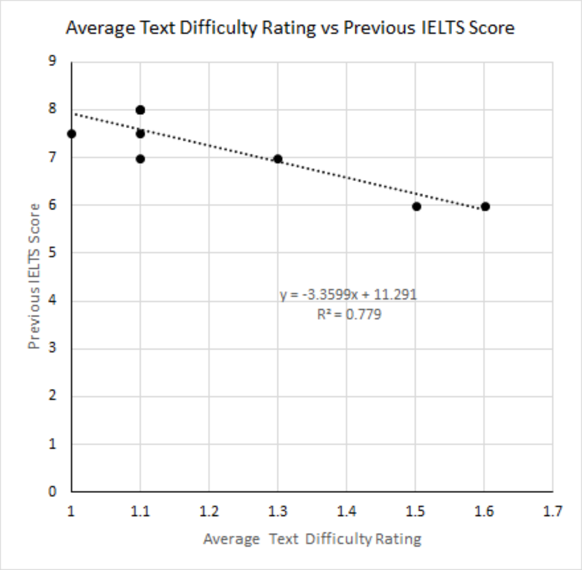

Figure 2: Average rating given by each participant versus their most recent IELTS score.

The only text to be given a rating above 3 by any participant was an extract from David Copperfield, which was presented as below.

> Mr. Micawber was extremely glad to see me, but a little confused too. He would have conducted me immediately into the presence of Uriah, but I declined.
>
> ’I know the house of old, you recollect,’ said I, ’and will find my way upstairs. How do you like the law, Mr. Micawber?’
>
> ’My dear Copperfield,’ he replied. ’To a man possessed of the higher imaginative powers, the objection to legal studies is the amount of detail which they involve. Even in our professional correspondence,’ said Mr. Micawber, glancing at some letters he was writing, ’the mind is not at liberty to soar to any exalted form of expression. Still, it is a great pursuit. A great pursuit!’

Of the public domain texts, The Woman in White (WW in Figure 1) was generally considered easier than the pool of classic literature by participants. Bookworms texts were generally considered easy, and David Copperfield was the most difficult.

6.2 Analysis of Text Features

On examining the relationship between the extracted features and the texts in each category (See Table 10), it was clear that the Bookworms text extracts had an almost monotonically increasing average sentence length from Level 0 (6.1) to 4 (13.95). All other texts had a higher average sentence length (15.6–22.5) except the YA text category (13.2).

Text Category	SPW	PDW	WPS
Children's lit.	1.26	17.7%	17.21
Level 0	1.19	21.9%	6.10
Level 1	1.19	22.8%	7.94
Level 2	1.20	16.2%	12.07
Level 3	1.23	15.5%	10.32
Level 4	1.31	24.9%	13.95
David Copperfield	1.32	22.5%	22.54
The Woman in White	1.29	18.4%	15.56
Top 3 classics	1.35	22.1%	20.33
Young Adult	1.29	24.7%	13.22

Table 10: Features averaged across each text category

While the average SPW increased monotonically for the Bookworm text levels, with all but Level 4 having a lower average SPW than other text categories, the PDW average varied considerably, with Level 4 having the highest average PDW across all categories of text. Clearly, the Dale-Chall word list is not a factor in setting the levels of Oxford Bookworm texts. Interestingly, the easiest non-Bookworm category based on average judgements (YA), had the second highest average PDW. A correlation across texts between SPW and WPS of 0.33 is probably due to the constraints placed on Oxford Bookworm text. While there was a fairly high correlation between SPW and PDW (0.53), clearly there were systematic differences.

7 Discussion of Results and Limitations

The models applied to the data did not display huge differences in effectiveness, but the Dale-Chall features generally outperformed the Flesch ones, in both the regression and SVM-based models. In the Flesch case, vocabulary became much more important as a feature relative to sentence length, compared to the original Flesch formula, whereas for Dale-Chall vocabulary was slightly less important than in the original formula.

Despite inconsistency in the relative importance of vocabulary to sentence length for Vietnamese speakers between feature sets, the collected text ratings showed a strong relationship with the level of English language skill of the participants, as measured by IELTS tests. On average across participants, text ratings appeared to follow logical trends, with Bookworm texts being easy, and classical literature being more challenging.

There were several limitations to our preliminary study, some of which can be addressed in future analysis, but most would require a new experiment with a greater number of participants. As with most empirical research, the more data available, the more robust the results. For any techniques that involve machine learning and many features, large sets of data are required. From a statistical perspective, having only about 100 data points only allows one to build good models involving multiple predictors if the effect size is expected to be large. We did, however, have a reasonable fit for both linear models on two features.

One of the difficulties was a mismatch between the participants and the experimental apparatus. Because of their relatively high level English background, being in the range IELTS 6–8 (for those who completed an IELTS test), the majority of the texts were too easy. This didn't provide enough discrimination between texts in the lower levels of difficulty. Based on an examination of the relationship between ratings and IELTS background, the current apparatus would require participants who have a much lower IELTS level.

There may have been a better model produced if the Likert scale was more fine-grained, to allow a greater spread of rating scores. For example, the research that developed the Lix readability model used a nine-point scale (Björnsson, 1968). However, one advantage of the Likert scale used here is that it should create more consistent ratings across participants, due to the precise wording for each point on the scale. Despite this, there is a drawback, in that the wording emphasises lexical difficulty. For a future study the wording should remove that emphasis to reduce potential bias toward vocabulary.

To make the test less onerous for lower level participants the wording would need to be further changed. Asking beginners to *read* a difficult text would result in them spending considerable time trying to decipher it, whereas what is required is a quick judgement as to its difficulty. So future questionnaires should ask participants to *look at* the text instead.

The most difficult text, as judged by Vietnamese participants, had long sentences and fairly long words, and a moderate percentage of difficult words, based on the Dale-Chall list. The texts judged the easiest had shorter sentences and words, and somewhat fewer difficult words. However, there was not a direct linear relationship between published simplified texts and human judgements. Perhaps there would be a linear relationship if the experiment had a different combination of participants and rating scale, but the current experiment does not provide evidence that the use of published scales as ground truth for human perception of reading difficulty is any more than a convenient substitute.

7.1 Future Work

We have been considering how to obtain a larger set of judgements from participants with lower levels of English comprehension, to provide a better spread of readability ratings. Using existing crowd-sourcing platforms is not really an option, since those who use them would already need a functional level of English to navigate the platforms. To our knowledge there is no equivalent platform available in Vietnamese.

We contemplated using students studying English in Vietnam, but this was unlikely to result in many participants due to the constraints on recruitment. An alternative may be using social media sites that are popular in Vietnam to advertise to participants, or providing a free Massive Open On-line Course (MOOC) for the collection of data from students. Evidence from this study and elsewhere (Jacob and Uitdenbogerd, 2019) suggests that to obtain enough participants with beginner or intermediate L2 skills, it is essential to recruit and present the study in their L1, or ratings will be exponential in distribution.

This initial study was limited to three traditional readability features. Further work would involve a wider range of features, with particular focus on those that are related to the human experience of reading (Crossley et al., 2008). However, a larger set of human judgements is needed before meaningful experimentation with ML techniques can be contemplated.

It may be useful to create a validated equivalent to the Dale-Chall list for Vietnamese speakers beyond our initial attempt at modifying the list with the input from a single Vietnamese participant. Due to the French colonial background of Vietnam there are also French-Vietnamese cognates (for example, ga tô for gateau) which may impact readability by being more memorable (Beinborn et al., 2014). However, the impact is likely to be much less than for more related language pairs such as Spanish-Italian, or French-English.

In this work we focused on the readability of passages of English text for speakers with Vietnamese L1. We are currently also working with other language backgrounds. Difficulty varies greatly across text, which with traditional formulae was managed by taking multiple samples. Flesch (1948) recommended 25-30 samples for measuring a book's readability, if the whole book is not being analysed, and Björnsson (1968) used 20 100-word samples for lexical complexity and 20 ten-sentence samples for grammatical complexity. We may explore sentence-level readability in future (Pilán et al., 2014).

8 Conclusion

This study is the first to attempt to measure English readability for Vietnamese speakers. Our contribution consists of a small data-set of human judgements of English text by Vietnamese volunteers, and the application of linear regression and SVM models to predict readability, using traditional readability features.

SVMs produced a model with the best performance in terms of MSE. Bootstrapping increased the MSE in linear models, but helped significantly in building an effective SVM model.

The features from Dale-Chall performed consistently well across all models (linear regression and SVMs). The small data-set prevented the rigorous use of a large feature set, despite a combination of statistical features and NLP features being likely to produce a better model (François and Miltsakaki, 2012). Thus future work includes applying more features to a larger data-set, preferably with a better match between text samples and participants, and using the Vietnamese data-set to tune a model produced from a larger data-set (Xia et al., 2016).

Acknowledgments

We thank Mr Patrick Jacob for suggesting different methodologies in the experiment.

References

Lisa Beinborn, Torsten Zesch, and Iryna Gurevych. 2014. Readability for foreign language learning: The importance of cognates. *ITL-International Journal of Applied Linguistics*, 165(2):136–162.

Carl-Hugo Björnsson. 1968. *Läsbarhet: hur skall man som författare nå fram till läsarna?* Bokförlaget Liber.

Scott A Crossley, Jerry Greenfield, and Danielle S McNamara. 2008. Assessing text readability using cognitively based indices. *Tesol Quarterly*, 42(3):475–493.

Edgar Dale and Jeanne S. Chall. 1948. A formula for predicting readability: Instructions. *Educational Research Bulletin*, 27(2):37–54.

Warwick B Elley and Francis Mangubhai. 1983. The impact of reading on second language learning. *Reading research quarterly*, pages 53–67.

R Flesch. 1948. A new readability yardstick. *Journal of Applied Psychology*, 32(3):221–233.

Thomas François and Eleni Miltsakaki. 2012. Do NLP and machine learning improve traditional readability formulas? In *Proceedings of the First Workshop on Predicting and Improving Text Readability for Target Reader Populations*, PITR '12, pages 49–57, Stroudsburg, PA, USA. Association for Computational Linguistics.

Thomas François. 2014. An analysis of a French as a foreign language corpus for readability assessment. In *Proceedings of the third workshop on NLP for computer-assisted language learning*, pages 13–32.

Frank Hermann. 2003. Differential effects of reading and memorization of paired associates on vocabulary acquisition in adult learners of english as a second language. *TESL-EJ*, 7(1):1–16.

Patrick Jacob and Alexandra L. Uitdenbogerd. 2019. In *Australasian Language Technology Workshop ALTW 2019*.

George R Klare. 1988. The formative years. In Beverly L. Zakaluk and S. Jay Samuels, editors, *Readability: Its past, present, and future*, pages 14–34. ERIC.

Batia Laufer. 1989. What percentage of text-lexis is essential for comprehension? In Christer Laurén and Marianne Nordman, editors, *Special language: From humans thinking to thinking machines*, pages 316–323.

Edward Loper and Steven Bird. 2002. NLTK: The natural language toolkit. In *Proceedings of the ACL-02 Workshop on Effective Tools and Methodologies for Teaching Natural Language Processing and Computational Linguistics - Volume 1*, ETMTNLP '02, pages 63–70, Stroudsburg, PA, USA. Association for Computational Linguistics.

John W Oller, J Donald Bowen, Ton That Dien, and Victor W Mason. 1972. Cloze tests in English, Thai, and Vietnamese: Native and non-native performance. *Language Learning*, 22(1):1–15.

Gustavo Paetzold and Lucia Specia. 2016. Semeval 2016 task 11: Complex word identification. In *Proceedings of the 10th International Workshop on Semantic Evaluation (SemEval-2016)*, pages 560–569.

F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay. 2011. Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12:2825–2830.

Ildikó Pilán, Elena Volodina, and Richard Johansson. 2014. Rule-based and machine learning approaches for second language sentence-level readability. In *Proceedings of the ninth workshop on innovative use of NLP for building educational applications*, pages 174–184.

Sarah E. Schwarm and Mari Ostendorf. 2005. Reading level assessment using support vector machines and statistical language models. In *Proceedings of the 43rd Annual Meeting on Association for Computational Linguistics*, ACL '05, pages 523–530, Stroudsburg, PA, USA. Association for Computational Linguistics.

A. L. Uitdenbogerd, S. Kablaoui, and A. Martin. 2017. Defining a unified model of vocabulary acquisition via extensive reading : Final report. Grant Report for the Office for Learning and Teaching.

A.L. Uitdenbogerd. 2005. Readability of French as a foreign language and its uses. *J. Kay, A. Turpin and R. Wilkinson (ed.) ADCS 2005: Proceedings of the Tenth Australasian Document Computing Symposium*, pages 19–25.

Sowmya Vajjala and Ivana Lucic. 2019. On understanding the relation between expert annotations of text readability and target reader comprehension. In *Proceedings of the Fourteenth Workshop on Innovative Use of NLP for Building Educational Applications*, pages 349–359, Florence, Italy. Association for Computational Linguistics.

Menglin Xia, Ekaterina Kochmar, and Ted Briscoe. 2016. Text readability assessment for second language learners. In *Proceedings of the 11th Workshop on Innovative Use of NLP for Building Educational Applications*, pages 12–22.